Named Entity Recognition in Speech to Text Transcripts

Peter Røysland Aarnes

Supervisors: Samia Touileb Lubos Steskal

Media Futures

Abstract

A UiB and TV2 collaboration project, aimed to ensure that all named entities appearing Norwegian speech-to-text transcripts are properly capitalized. Using various machine learning methods, aimed at accurately identify named entities where proper capitalization is not already in place.

To validate the models used for proper capitalization, their predictions be compared to professional human transcripts that will be acting as a gold standard provided by TV2.

Research question

1. What method will produce the most accurate results, neural models such as a BiLSTM or Norwegian based BERT models?

2. How would a binary annotation scheme in the effect the results compared to a IOB2 annotation format when training the models?

3. What method is the best in terms of speed and computing efficiency?

Why is Named Entity Recognition difficult?

Sentences can have vastly different meanings, even if the spelling is the same:

«Jobs said…»
«Jobs are hard to find»

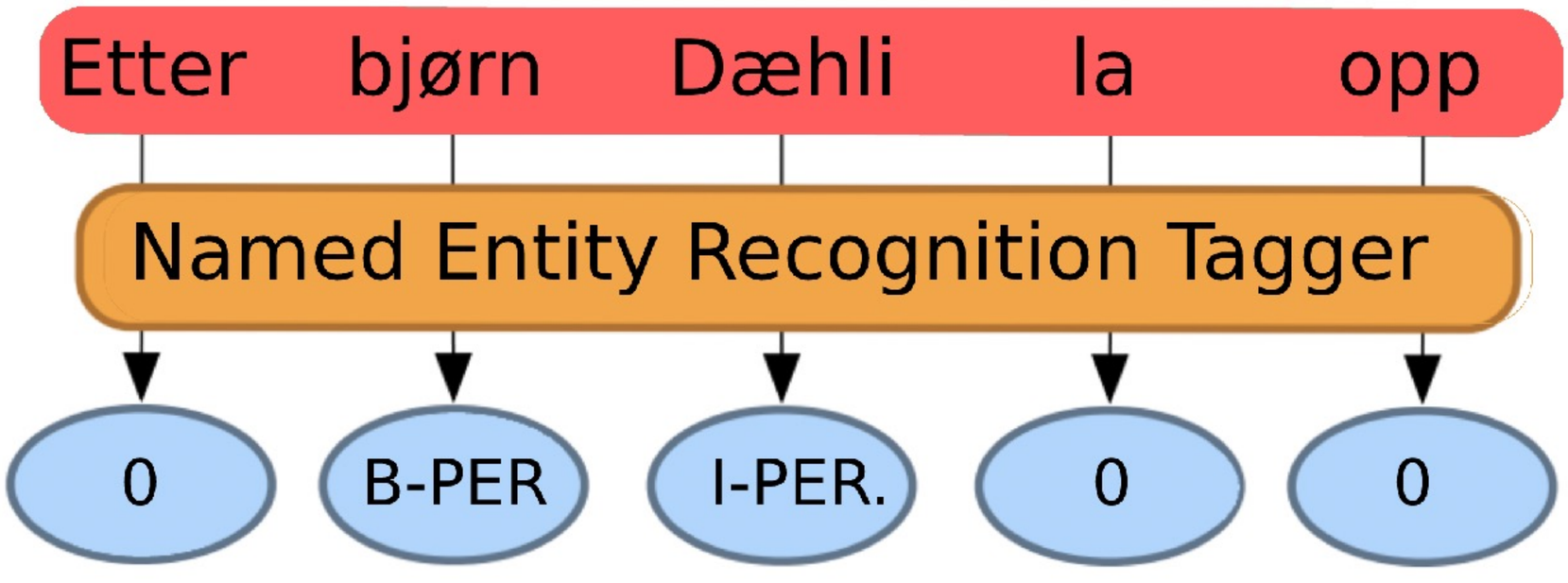
Relying on gazetteers, large knowledge bases for named entities to identify named entities would not be feasible because the open nature of names.

A lack of sufficient training data could also pose a problem.
- Available richly annotated training data in Norwegian is sparse.
- NorNE corpus, 600 000 tokens

Possible ways to extent training corpus
- Using enormous amounts of text and automatically annotate capitalized words in a binary fashion.

Research contribution

- Expanding knowledge for named entity research and natural language research where text has improper capitalization
- Provide TV2 with models that could improve their speech-to-text technology



PARTNERS

HOST

UNIVERSITY OF BERGEN

FUNDER

This research is funded by SFI MediaFutures partners and the Research Council of Norway (grant number 309339).

Forskningsrådet